

02.06.00

JP00/3625

4
日本特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application:

1999年 6月 4日

27 JUL 2000

出願番号
Application Number:

平成11年特許願第158498号

出願人
Applicant(s):

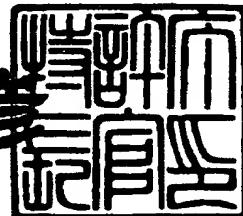
セイコーエプソン株式会社

PRIORITY
DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

2000年 6月 29日

特許庁長官
Commissioner,
Patent Office

近藤 隆彦



出証番号 出証特2000-3052006

【書類名】 特許願
【整理番号】 J0074155
【提出日】 平成11年 6月 4日
【あて先】 特許庁長官 殿
【国際特許分類】 G06F 17/30
G06F 7/36
【発明の名称】 文書分類方法及び文書分類装置並びに文書分類処理プロ
グラムを記録した記録媒体
【請求項の数】 12
【発明者】
【住所又は居所】 長野県諏訪市大和3丁目3番5号 セイコーエプソン株
式会社内
【氏名】 長石 道博
【特許出願人】
【識別番号】 000002369
【氏名又は名称】 セイコーエプソン株式会社
【代表者】 安川 英昭
【代理人】
【識別番号】 100093388
【弁理士】
【氏名又は名称】 鈴木 喜三郎
【連絡先】 0266-52-3139
【選任した代理人】
【識別番号】 100095728
【弁理士】
【氏名又は名称】 上柳 雅誉
【選任した代理人】
【識別番号】 100107261
【弁理士】

【氏名又は名称】 須澤 修

【手数料の表示】

【予納台帳番号】 013044

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9711684

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書分類方法及び文書分類装置並びに文書分類処理プログラムを記録した記録媒体

【特許請求の範囲】

【請求項1】 複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行い、

このクラスタマージ処理によって得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、その表示内容をユーザに提示すべき分類結果に含めて出力することを特徴とする文書分類方法。

【請求項2】 前記クラスタマージ処理内容がわかるような表示内容とは、前記統合されたそれぞれのクラスタ間の関連性の高さに基づき、当該それぞれのクラスタのクラスタ名の表示の仕方を変えた表示内容であって、それぞれのクラスタ名の表示の仕方は、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をAND形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さが予め設定された値未満である場合には、それぞれのクラスタ名をOR形式の表記の仕方で表示させることを特徴とする請求項1記載の文書分類方法。

【請求項3】 前記AND形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名を横方向に並べて連続的に表記するか、それぞれのクラスタ対応のクラスタ名ごとに改行して縦に並べて表記するかのいずれかで行い、前記OR形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名の間に区切り記号を挿入して表記することを特徴とする請求項2記載の文書分類方法。

【請求項4】 あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することを特徴とする請求項2または3に記載の文書分類方法。

【請求項5】 複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、

このクラスタリング部によって得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部と、

このクラスタマージ部によってクラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ内容生成部と、

その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力手段と、

を有したことを特徴とする文書分類装置。

【請求項6】 前記前記クラスタマージ処理内容生成手段が生成するクラスタマージ処理内容がわかるような表示内容は、前記統合されたそれぞれのクラスタ間の関連性の高さに基づき、当該それぞれのクラスタのクラスタ名の表示の仕方を変えた表示であって、それぞれのクラスタ名の表示の仕方は、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をAND形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さが予め設定された値未満である場合には、それぞれのクラスタ名をOR形式の表記の仕方で表示させることを特徴とする請求項5記載の文書分類装置。

【請求項7】 前記AND形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名を横方向に並べて連続的に表記するか、それぞれのクラスタ対応のクラスタ名ごとに改行して縦に並べて表記するかのいずれかで行い、前記OR形式

の表記の仕方は、それぞれのクラスタ対応のクラスタ名の間に区切り記号を挿入して表記することを特徴とする請求項6記載の文書分類装置。

【請求項8】 あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することを特徴とする請求項6または7に記載の文書分類装置。

【請求項9】 複数の文書を意味的に共通性を有する複数のクラスタに分類して出力する文書分類処理プログラムを記録した記録媒体であって、その処理プログラムは、

複数の文書を意味的に共通性を有する複数のクラスタに分類する手順と、
その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行う手順と、

クラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成する手順と、

その表示内容をユーザに提示すべき分類結果に含めて出力する手順と、
を含むことを特徴とする文書分類処理プログラムを記録した記録媒体。

【請求項10】 前記クラスタマージ処理内容がわかるような表示内容とは、前記統合されたそれぞれのクラスタ間の関連性の高さに基づき、当該それぞれのクラスタのクラスタ名の表示の仕方を変えた表示内容であって、それぞれのクラスタ名の表示の仕方は、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をAND形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さが予め設定された値未満である場合には、それぞれのクラスタ名をOR形式の表記の仕方で表示させることを特徴とする請求項9記載の文書分類処理プログラムを記録した記録媒体。

【請求項11】 前記AND形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名を横方向に並べて連続的に表記するか、それぞれのクラスタ対応のクラスタ名ごとに改行して縦に並べて表記するかのいずれかで行い、前記OR形

式の表記の仕方は、それぞれのクラスタ対応のクラスタ名の間に区切り記号を挿入して表記することを特徴とする請求項10記載の文書分類処理プログラムを記録した記録媒体。

【請求項12】 あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することを特徴とする請求項10または11に記載の文書分類処理プログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は多数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体に関する。

【0002】

【従来の技術】

多数の文書を意味的なまとまりごとの複数のクラスタに分類する際、それぞれの文書から特徴要素を抽出し、その特徴要素に基づいて分類することが行われている。その分類手法として、それぞれの文書全体（表題や本文など1つの文書を構成する文書内容全体）を特徴要素の抽出対象とし、それぞれの文書全体から特徴要素を抽出し、抽出された特徴要素に基づいて複数のクラスタに分類する文書分類方法がある。

【0003】

この文書全体を特徴要素抽出の対象として分類を行うと、文書の形態素解析や、特徴抽出処理が非常に繁雑であり、CPUがその処理を行う場合、CPUに対する負荷を大きいものとしている。また、一般に、文書はその文書の主旨とは直接関係のない記述を多く含んでいるのが普通である。したがって、文書全体を特徴要素抽出の対象とすると、それによって分類されたクラスタは情報の分類という観点から見たとき、あまり意味のない分類となることが多い。つまり、ノイズクラスタが多数生成されてしまうということにもなる。

【0004】

このような問題点を解決する手法として、それぞれの文書の主旨を適切に表す部分としてそれぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法がある。この手法は、文書の主旨を適切に反映した文書分類を可能とすることができるものとして期待される。

【0005】

このように、従来から文書を幾つかのクラスタに分類する手法は幾つか考えられている。

【0006】

【発明が解決しようとする課題】

しかしながら、上述した適切な分類がなされる手法である文書の表題部から抽出された特徴要素に基づいて文書を分類する手法を用いたとしても、それによって得られる分類結果は、クラスタの数が多くなりすぎることもあり、ユーザ側から見たときに、決して適切な分類が行われたとは思えない場合もでてくる。たとえば、分類されて得られる多数のクラスタを比較した場合、それぞれのクラスタに共通した文書が数多く含まれる場合もある。このような場合、ユーザは提示された多数のクラスタについて、結局は、自分で整理し、その中から自分の本当に欲しい情報を探すというような面倒な処理を行うことになる。

【0007】

これに対処する方法として、分類されて得られた多数のクラスタに対し、それぞれのクラスタ間で内容の関連性を判断し、関連性のあるクラスタ同志を統合して新たなクラスタを生成するクラスタマージ技術も提案されている。

【0008】

このようなクラスタマージ処理を行うと、最初のクラスタリング処理によって、多数のクラスタが生成されたとしても、内容の関連性の高いクラスタ同志が統合されるので、簡潔化されたクラスタリング結果をユーザに提示することができ、ユーザは自分の欲しい情報を効率よく探すことができるようになる。

【0009】

しかし、クラスタマージ後のクラスタをユーザに提示する際、単に、クラスタマージ処理結果が提示されたとすると、ユーザ側からみたとき、どのようなクラスタマージ処理がなされて統合されたのかといったクラスタマージ処理内容、すなわち、そのクラスタマージによって得られた新たなクラスタは、もともとどのクラスタとどのクラスタがどの程度の関連性があるから統合されたのかといった内容がわかりにくくなることがある。

【0010】

そこで、本発明は、内容に関連性のある複数のクラスタを統合するクラスタマージ処理がなされたあと、そのクラスタマージ処理されて得られたら新たなクラスタを表示する際、その新たなクラスタは、どのクラスタとどのクラスタがどの程度の関連性があるから統合されたのかといったクラスタマージ処理内容がわかるように表示することを目的としている。

【0011】

【課題を解決するための手段】

前述の目的を達成するために、本発明の文書分類方法は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行い、このクラスタマージ処理によって得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、その表示内容をユーザに提示すべき分類結果に含めて出力するようにしている。

【0012】

また、本発明の文書分類装置は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、このクラスタリング部に

よって得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部と、このクラスタマージ部によってクラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ内容生成部と、その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力手段とを有した構成としている。

【0013】

さらに、本発明の文書分類処理プログラムを記録した記録媒体は、複数の文書を意味的に共通性を有する複数のクラスタに分類して出力する文書分類処理プログラムを記録した記録媒体であって、その処理プログラムは、複数の文書を意味的に共通性を有する複数のクラスタに分類する手順と、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行う手順と、クラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成する手順と、その表示内容をユーザに提示すべき分類結果に含めて出力する手順とを含むようにしている。

【0014】

これら各発明において、前記クラスタマージ処理内容がわかるような表示内容とは、前記統合されたそれぞれのクラスタ間の関連性の高さに基づき、当該それぞれのクラスタのクラスタ名の表示の仕方を変えた表示内容であって、それぞれのクラスタ名の表示の仕方は、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をAND形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さが予め設定された値未満である場合には、それぞれのクラスタ名をOR形式の表記の仕方で表示させるようにしている。

【0015】

そして、前記AND形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名を横方向に並べて連続的に表記するか、それぞれのクラスタ対応のクラスタ名ごとに改行して縦に並べて表記するかのいずれかで行い、前記OR形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名の間に区切り記号を挿入して表記するようにしている。

【0016】

さらに、あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することも可能としている。

【0017】

このように本発明は、クラスタマージされて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、それを表示するようにしている。

【0018】

これによって、ユーザは、クラスタマージされる前のクラスタの様子、すなわち、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかといったことを表示内容を見るだけで知ることができる。そして、どのような関連性を有しているかを示す表示の仕方としては、クラスタマージ処理されて得られた新たなクラスタに含まれるクラスタ間の関連性の高さに基づき、クラスタマージ処理されたそれぞれのクラスタのクラスタ名の表示の仕方を変えるようにしている。

【0019】

そのクラスタ名の表示の仕方は、具体的には、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をAND形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さを表す値が予め設定された値未満である場合には、それぞれのクラスタ名をOR形式の表記の仕方で表示させるようにしている。たとえば、関連性の高さがきわめて高い場合には、それ

そのクラスタ名を横一列に連続的に表示したり、それぞれのクラスタ名を1つづつ縦に並べて表示し、関連性の高さがそれほどでもない場合には、それぞれのクラスタ名の間に区切り記号を挿入するなどして表示する。ユーザはこのような表示を見ることで、統合される前のそれぞれのクラスタがどのようなクラスタであって、それぞれのクラスタ同志の関連性がどの程度であるかなどを知ることができる。

【0020】

また、あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することも可能であり、包含関係であることを繁雑なイメージを使わないでもわかりやすく表示できる。

【0021】

【発明の実施の形態】

以下、本発明の実施の形態について説明する。なお、この実施の形態で説明する内容は、本発明の文書分類方法および文書分類装置についての説明であるとともに、本発明の文書分類処理プログラムを記録した記録媒体における文書分類処理プログラムの具体的な処理内容をも含むものである。

【0022】

また、この実施の形態では、文書分類の手法として、前述したように、それぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法を用いるものとする。

【0023】

図1は本発明を実現するための装置構成を示すもので、大きく分けると、それぞれの文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部1と、このクラスタリング部1によって得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部2と、このクラスタマージ部2によってクラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内

容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ処理内容生成部3と、その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力部4とを有した構成となっている。

【0024】

クラスタリング部1は、文書記憶部11、文解析部12、特徴要素抽出部13、特徴テーブル作成部14、文書分類部15、分類結果記憶部16を有している。

【0025】

文書記憶部11はこの場合、多数の文書データをデータベースとして持つものである。ここでは、たとえば、図2に示すような文書群を分類する場合を説明する。図2に示される文書群は、それぞれが独立した文書D1, D2, ..., D7を有し、これらの文書D1, D2, ..., D7は表題部T1, T2, ..., T7と、それに対する本文A1, A2, ..., A7を持っているものとする。

【0026】

文解析部12は文書記憶部11に記憶されている文書を文解析し、それぞれの文書の表題部を検出する。この文解析部12が行う表題部の検出は、具体的には次のようにして行う。

【0027】

まず、第1の方法として、文書構造様式によって表題と規定される部分があればその部分を表題部とする。また、第2の方法として、文書構造様式によって、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とする。また、第3の方法として、定められた数の文または単語を文書先頭より抽出し、その抽出した部分を表題部とする。さらには、これら第1、第2、第3の方法を順次行い、第1の方法を行ったとき、表題と規定されている部分があればその部分を表題部とし、表題と規定される部分が存在しなければ、第2の方法を行い、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とし、標準より大きな文字で表示する指定がなされていなければ

ば、第3の方法を行って表題部を検出する。

【0028】

特徴要素抽出部13は、文解析部2で検出されたそれぞれの文書の表題部の中から特徴要素を抽出する。

【0029】

特徴テーブル作成手段14は、前記表題部から抽出された特徴要素とそれぞれの文書との関係を示す特徴テーブルを作成する。なお、この特徴テーブルの具体的な内容については後述する。

【0030】

文書分類部15は、前述の特徴テーブルの内容を参照し、文書D1, D2, ···, D7を意味的に共通性のある複数のクラスタに分類する。つまり、文書D1, D2, ···, D7の表題部に存在する特徴要素に基づいて、共通する特徴要素を持つ処理対象文書を1つのまとまりとし、そのまとまりを1つのクラスタとする。なお、この文書分類部15は同義特徴辞書（図示せず）を有し、共通する特徴要素を持つ処理対象文書を1つのまとまりとする処理を行う際、共通する特徴要素であるか否かの判断を、その同義語辞書を用い同義語が有るか否かにより行い、同義語が存在する場合にはそれを同じクラスタとする処理を行うことも可能である。

【0031】

分類結果記憶部16は、文書分類部15によって分類された内容を記憶する。

【0032】

クラスタマージ部2は、複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合する処理を行うものであるが、その具体的な処理については後述する。

【0033】

クラスタマージ処理内容生成部3は、前記クラスタマージ部2で判断されたクラスタ間の関連性の高さを示す値（後述する）を用い、その値を予め設定されたしきい値（後述する）と比較して関連性の高さを判断する関連性判断部31と、

この関連性判断部31によるクラスタ間の関連性の高さに基づいて、どのようなクラスタがどのような関連性を有して統合されたかがわかるように、それぞれのクラスタ名の表示の仕方を決めるクラスタ名表示内容決定部32とを有し、その具体的な処理内容については後述する。

【0034】

また、分類結果出力部4は、出力制御部41と表示部42を有し、本発明による文書分類結果を出力する。

【0035】

このような構成において、本発明の文書分類処理について説明する。本発明が行う概略的な文書分類処理は、図3のフローチャートに示すように、処理対象となる多数の文書を意味的に共通性を有する複数のクラスタに分類し（ステップS1）、これにより分類された複数のクラスタ間で各々のクラスタに含まれる文書に基づいて、それぞれのクラスタ間の関連性を判断する（ステップS2）。そして、一定以上の関連性を有する少なくとも2つのクラスタを統合する（ステップS3）。その後、クラスタマージされて得られた新たなクラスタは、どのようなクラスタがどのような関連性を有して統合されたかがわかるようなクラスタマージ内容を生成する。具体的には、クラスタマージされたクラスタ間の関連性の高さを判定し（ステップS4）、その関連性の高さに基づいて、統合される前の個々のクラスタに関する情報がわかるような表示内容、すなわち、クラスタマージによって得られた新たなクラスタは、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかがわかるような表示内容を生成する（ステップS5）。以下、具体例を参照して詳細に説明する。

【0036】

ここでは、図2で示した文書D1, D2, ..., D7を分類する例について説明する。この実施の形態では、それぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理された結果についてクラスタマージ処理を行う。まず始めに、表題部から特徴要素を抽出し、その抽出された特徴要素に基づいて行われるクラスタリング処理（クラスタリング部1が行う処理）について説明する。

【0037】

これらの文書D1, D2, …, D7は、文解析部12にて表題部が検出される。たとえば、文書D1については表題部T1が検出され、文書D2については表題部T2が検出され、文書D3については表題部T3が検出されるというよう、それぞれの文書D1, D2, …, D7の表題部T1, T2, …, T7が検出される。

【0038】

そして、特徴要素抽出部13によって、それぞれの表題部に存在する特徴要素が抽出されたのち、特徴テーブル作成部14により、それぞれの特徴要素とその特徴要素を表題部に含む文書との関係を示す特徴テーブルが作成される。この特徴テーブルの例を図4に示す。なお、ここでは、文書数が3つ以上取り出される特徴要素とその特徴要素を含む文書との関係を示し、特徴テーブル内に示される数値は、その特徴要素が各文書の表題部に幾つ含まれているかの数を示している。たとえば、「用紙」という特徴要素は、文書D1, D4, D6, D7のそれぞれの表題部に、それぞれ1個ずつ含まれていることを示している。

【0039】

図4の特徴テーブルからもわかるように、表題部に「用紙」という特徴要素を含む文書は、文書D1, D4, D6, D7であり、また、表題部に「カセット」という特徴要素を含む文書は、文書D1, D4, D7であり、さらに、表題部に「増設」という特徴要素を含む文書は、文書D2, D3, D5, D7である。なお、図2において、これら各特徴要素部分にはアンダーラインが施されている。

【0040】

そして、文書分類部15はこのような特徴テーブルを参照して、それぞれの特徴要素ごとの文書クラスタ分けを行う。その分類結果を図5に示す。なお、このようなクラスタに分類する際、前述したように、共通する特徴要素であるか否かの判断を、同義語辞書を用い同義語が有るか否かによっても行い、同義語が存在する場合にはそれを同じ文書クラスタとする処理を行うことも可能である。たとえば、「用紙」と「印刷紙」の両方が特徴要素として抽出されたとすれば、これらの特徴要素を表題部に含む文書は同じクラスタとするなどという処理を行う。

【0041】

このような分類結果は分類結果記憶部16に格納される。図5に示される分類結果において、たとえば、「用紙」で分類されたクラスタ（文書D1, D4, D6, D7が含まれる）について見れば、図2の文書内容からもわかるように、文書D1は用紙カセットについての内容であり、文書D4は用紙設定についての内容であり、文書D6は印刷された後の用紙の汚れについての内容であり、文書D7は用紙カセットの増設についての内容である。

【0042】

このように、これらの文書D1, D4, D6, D7はどれも用紙に関する内容であり、1つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

【0043】

また、「カセット」で分類されたクラスタ（文書D1, D4, D7が含まれる）について見れば、図2の文書内容からもわかるように、文書D1は用紙カセットについての内容であり、文書D4は用紙設定についての内容であり、文書D7は用紙カセットの増設についての内容である。

【0044】

このように、これらの文書D1, D4, D6, D7にはどれも用紙をセットすることに関する内容が含まれており、1つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

【0045】

また、「増設」で分類されたクラスタ（文書D2, D3, D5, D7が含まれる）について見れば、図2の文書内容からもわかるように、文書D2はメモリの増設についての内容であり、文書D3はインターフェースカードの増設についての内容であり、文書D5はハードディスクの増設についての内容であり、文書D7は用紙カセットの増設についての内容である。

【0046】

このように、これらの文書D2, D3, D5, D7はどれも何かを増設する場合についての内容であり、1つのクラスタとして分類されて何等問題のないもの

となり、その分類結果は適切であるといえる。

【0047】

このような適切な分類が行える理由としては、それぞれの文書の表題部から特徴要素を抽出し、その特徴要素に基づいて文書を分類しているからである。つまり、文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることが多い。したがって、文書の表題部に含まれる特徴要素を用いて分類を行うことにより、分類結果が散漫になることが少なく、また、ノイズクラスタが生成される率も少なくすることができる。また、各文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることから、文書の制作者側の視点による分類が得られる。

【0048】

そして、分類が行われた後、ユーザによって、たとえば、「用紙」についてのクラスタの選択指示が出されたとすると、そのクラスタに属する文書D1, D4, D6, D7が文書記憶部11から読み出されて表示部32に表示される。なお、このときの表示内容としては、前述したように、文書番号や文書名のみでもよく、さらには、その文書内容を表示させるようにしてもよい。

【0049】

ところで、本発明は以上のようにクラスタリング処理した結果について、さらに、クラスタマージ部2によってクラスタマージ処理を行う。

【0050】

すなわち、図5に示す分類結果において、特徴要素である「用紙」と「カセット」について見ると、「用紙」のクラスタには文書D1, D4, D6, D7が含まれ、「カセット」のクラスタには文書D1, D4, D7に存在することがわかる。

【0051】

このように、「用紙」のクラスタと「カセット」のクラスタには、共に文書D1, D4, D7が共通して存在している。これは、「用紙」という特徴要素と「カセット」という特徴要素は相互に関連した状態で用いられることが多いことを意味している。たとえば、文書D1, D4, D7の表題部または本文のなかに「

用紙力セット」という用語が用いられている。つまり、これらの文書D1, D4, D7は共通性の高い文書であり、これら文書D1, D4, D7は同じクラスタに分類した方がより好ましいと考えられる。

【0052】

これを実現するために、特徴要素に基づいてクラスタリングしたあと、そのクラスタリング結果に対しクラスタマージ処理を施す。

【0053】

このクラスタマージ処理について以下に説明する。まず始めに、図5の分類結果とは関係なく一般的な例について図6を参照しながら説明する。

【0054】

今、2つのクラスタC1, C2があるとする。クラスタC1として5個の文書D1, D2, D3, D4, D8が抽出され、クラスタC2には6個の文書D3, D4, D5, D6, D7, D8が抽出されたとする。

【0055】

ここで、2つのクラスタC1, C2に共通している文書は、文書D3, D4, D8である。この実施の形態では、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基に、それぞれのクラスタ間の関連性を判断してクラスタマージ処理を行う。

【0056】

具体的には、複数のクラスタとして、ある2つのクラスタに共通している文書数が2つのクラスタに存在する合計の文書数に対しどのくらいの割合かを計算し、その計算結果が予め定めたしきい値以上かどうかによってマージするか否かを決める。

【0057】

たとえば、この場合、2つのクラスタC1, C2に存在する文書数の合計は11個であり、両者に共通する文書数は3個である。これらから合計の文書数に占める共通する文書数の割合(%)を計算し、その結果からマージするか否かを決定する。この割合(%)を求める際、合計の文書数で共通する文書数を単純に割り算してそれに100を掛けて求めてよいが、共通する文書数に任意に設定さ

れる係数を掛け算したものを合計の文書数で割り算してそれに100を掛けて求めるようにしてもよい。

【0058】

一例として、クラスタC1に存在する文書数を α_1 、クラスタC2に存在する文書数を α_2 とし、両者に共通する文書数を β とした場合、たとえば β に係数としてたとえば2を掛けて、 $2\beta / (\alpha_1 + \alpha_2) \times 100$ を計算し、その値(%)が予め設定されたしきい値TH(%)と比較して、上式による計算結果がしきい値TH以上であればマージするというようなことを行う。図6で示した例について考えれば、 2β は $2 \times 3 = 6$ 個、 $\alpha_1 + \alpha_2$ は $5 + 6 = 11$ 個であるので、この場合、約55%と求められる。ここで、しきい値THが仮に70%と設定されているとすれば、計算結果(55%)はしきい値TH(70%)より小さいので、クラスタC1とクラスタC2はマージしないとする。なお、係数は任意に設定されるもので、計算結果で得られる数値(%)がしきい値と比較し易いような値となるように適当に設定されるものであり、この場合は係数を2としたが、係数を1としても特に問題はない。

【0059】

ここで、図5で示した分類結果を例にして説明すれば、図5の場合、「用紙」のクラスタには文書D1, D4, D6, D7の4つの文書が存在し、「カセット」のクラスタには文書D1, D4, D7の3つの文書が存在する。そして、2つのクラスタに共通する文書は文書D1, D4, D7の3つの文書であり、これを合計の文書数に対する割合(%)で考える。

【0060】

これを前述した計算式によって計算する。図5の分類結果の場合、合計の文書数($\alpha_1 + \alpha_2$)は、 $4 + 3 = 7$ となり、共通の文書数は3で 2β は6となる。したがって、この場合、約86%という高い値が得られる。これは、設定されたしきい値(ここでは70%としている)よりも高いので、この「用紙」のクラスタと「カセット」のクラスタはマージして1つのクラスタとするということになる。

【0061】

同様に考えて、図5の「用紙」のクラスタと「増設」のクラスタとをマージするか否か、「カセット」のクラスタと「増設」のクラスタとをマージするか否かについて判断する。

【0062】

まず、「用紙」のクラスタと「増設」のクラスタについては、「用紙」のクラスタには文書D1, D4, D6, D7の4つの文書が存在し、「増設」のクラスタには文書D2, D3, D5, D7の4つの文書が存在する。そして、2つのクラスタに共通する文書は文書D7のみであり、これを上式を用いて計算すると、この場合、25%という結果が得られ、これは、しきい値(70%)よりも低い値であるので、この場合は、両者はマージしないとする。

【0063】

また、「カセット」のクラスタと「増設」のクラスタについては、「カセット」のクラスタには文書D1, D4, D7の3つの文書が存在し、「増設」のクラスタには文書D2, D3, D5, D7の4つの文書が存在する。そして、2つのクラスタに共通する文書は文書D7のみであり、これを上式を用いて計算すると、この場合、約28%という結果が得られ、これは、しきい値(70%)よりも低い値であるので、この場合は、両者はマージしないとする。

【0064】

このようにして、それぞれのクラスタに対し2つのクラスタごとにそれぞれマージするか否かを判断する。この図5の分類結果についてマージするか否かの処理を行ったあとの分類結果(マージ処理後の分類結果という)が図7である。図7によれば、「用紙」と「カセット」が「用紙+カセット」という1つのクラスタに分類され、そのクラスタに属する文書は文書D1, D4, D6, D7ということになる。また、「増設」についてはそのまま単独のクラスタを構成する。

【0065】

図7に示されるクラスタマージ処理後の分類結果において、たとえば、「用紙+カセット」で分類されたクラスタ(文書D1, D4, D6, D7が含まれる)について見れば、図2の文書内容からもわかるように、文書D1は用紙カセット

についての内容であり、文書D4は用紙設定についての内容であり、文書D6は印刷された後の用紙の汚れた場合にはどのようにするかについての内容であり、文書D7は用紙カセットの増設についての内容である。

【0066】

このように、これらの文書D1, D4, D6, D7はどれも用紙やカセットに関する内容であり、1つのクラスタとして分類されて何等問題のないものとなり、むしろ、「用紙+カセット」を1つのクラスタとした方がよい分類結果であるといえる。

【0067】

このように、始めにそれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理されて得られたそれぞれのクラスタに対し、2つづつのクラスタの組み合わせについてクラスタマージ処理を行うことによって、より適切なクラスタリングが行える。

【0068】

また、以上のようにして2つのクラスタごとに1回目のクラスタマージ処理が終了し、図7のようなクラスタマージ処理後の分類結果が得られると、今度は、そのクラスタマージ処理後の分類結果について、2回目のクラスタマージ処理を行う。つまり、図7の1回目のクラスタマージ処理後の結果で考えた場合、「用紙+カセット」のクラスタと「増設」のクラスタについてクラスタマージ処理を行う。この場合、「用紙+カセット」のクラスタと「増設」のクラスタについては、「用紙+カセット」のクラスタには文書D1, D4, D6, D7の4つの文書が存在し、「増設」のクラスタには文書D2, D3, D5, D7の4つの文書が存在する。そして、2つのクラスタに共通する文書は文書D7のみであり、これを合計の文書数に対する割合(%)で考えると、共通する文書数1に定数2を掛けたものを合計の文書数8で割り算し、それに100を掛けると、25%という結果が得られ、これは、しきい値(70%)よりも低い値であるので、この場合は、両者はマージしないとする。

【0069】

このようにして、2つのクラスタ間で1回目のクラスタマージ処理が終了した後、その1回目のクラスタマージ処理に新たな2つのクラスタ間で2回目のクラスタマージ処理を行い、その2回目のクラスタマージ処理が終了した後、その2回目のクラスタマージ処理後に新たな2つのクラスタ間で3回目のクラスタマージ処理を行うというクラスタマージ処理を順次行い、新たなクラスタが生成されなくなるまで（クラスタマージが起こらなくなるまで）その処理を繰り返す。

【0070】

また、これまでの説明では、2つのクラスタ間でクラスタマージ処理を行う例について説明したが、クラスタマージ処理は3つ以上のクラスタの組み合わせについても可能である。この場合、1回のクラスタマージ処理によって3つ以上のクラスタ間でクラスタマージ処理を行い、さらに、これによって幾つかのクラスタに分類された結果についてクラスタマージが起こらなくなるまで、順次、クラスタマージ処理を行うことも可能である。なお、3つ以上のクラスタについてクラスタマージするか否かを判断する場合、前述したように、それぞれのクラスタに存在する合計の文書数に対する共通の文書数の割合（%）で考えることができる。

【0071】

以上のようにしてクラスタマージ部2によるクラスタマージ処理が終了すると、次に、クラスタマージ処理内容生成部3がそのクラスタマージ結果に対し、クラスタマージされたクラスタ間の関連性の高さを判定し、その関連性の高さに基づいて、統合される前の個々のクラスタに関する情報がわかるような表示内容、すなわち、クラスタマージによって得られた新たなクラスタは、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかがわかるような表示内容を生成する。以下、このクラスタマージ処理内容生成部3が行う処理について説明する。

【0072】

この実施の形態では、クラスタマージ部2によって得られた関連性の高さとしての関連性の度合い（%）の値が、前述したしきい値THよりずっと大きい値で

あるか、しきい値THに近い値であるかによって、そのクラスタマージされたクラスタ間の関連性の高さを関連性判断部31によって判断する。具体的には、前述のしきい値THに対し、それよりも高い値(%)のしきい値TH1を設定し、クラスタマージ部2によって得られた関連性の度合い(Kで表す)が、 $K \geq TH_1$ であれば、クラスタ同志の関連性はきわめて大きく殆ど同じ内容であると判断する。一方、クラスタマージ部2によって得られた関連性の度合いKが、 $TH_1 > K \geq TH$ であれば、少し似ている程度と判断する。

【0073】

今、 $K \geq TH_1$ である場合、すなわち、クラスタマージされて得られた新たなクラスタに含まれる幾つかのクラスタ同志の関連性がきわめて高い場合は、次のような処理を行う。

【0074】

これを図7の例で説明すれば、クラスタマージされた新たなクラスタの特徴要素は、「用紙+カセット」である。この「用紙+カセット」のクラスタは、図5に示す用紙のクラスタとカセットのクラスタをクラスタマージした結果である。このそれぞれのクラスタにクラスタ名を付けるとすれば、特徴要素が「用紙」であるクラスタを「用紙クラスタ」、特徴要素が「カセット」であるクラスタを「カセットクラスタ」といように表すことができ、それぞれのクラスタ名を以下では、単に、「用紙」、「カセット」と表記する。

【0075】

ここで、クラスタマージ結果である「用紙+カセット」のクラスタは、クラスタマージ部2による前述の計算によって、86%という値が得られている。ここで、関連性判断部31において、関連性を判断する際に設定されたしきい値TH1が80%と設定されているとすれば、この場合、クラスタマージ部2によって得られた関連性の度合いKは、 $K \geq TH_1$ であるので、用紙クラスタとカセットクラスタの関連性はきわめて大きく殆ど同じ内容であると判断できる。

【0076】

このように、クラスタマージ部2によって得られた関連性を示す値Kが、 $K \geq TH_1$ である場合には、クラスタマージされたそれぞれのクラスタ同志の関連性

はきわめて大きく、殆ど同じ内容のクラスタであると判断でき、それぞれのクラスタの名称を、連続的に表示する。たとえば、上述の「用紙クラスタ」と「カセットクラスタ」の例では、それらのクラスタ名である「用紙」と「カセット」をくっつけて「用紙カセット」などと表記してそれを表示する。

【0077】

これは、いわゆるAND形式の表記の仕方であり、クラスタ名をくっつけて表記しても差し支えないような場合である。この例では、クラスタマージされて得られた新たなクラスタのクラスタ名を「用紙カセット」とすることになるが、この場合は、クラスタマージされて得られた新たなクラスタは、その新たなクラスタを構成する用紙クラスタとカセットクラスタに含まれるそれぞれの文書内容（図2参照）から見て、新たなクラスタ名を「用紙カセット」として何等差し支えないものである。

【0078】

図8はこのような処理を行ったあとの表示例を示すもので、この図8では、クラスタマージされた新たなクラスタのクラスタ名としての「用紙カセット」と、その新たなクラスタに含まれる文書として、ここでは、図2で示されたそれぞれの文書（文書D1, D4, D6, D7）のそれぞれの表題（タイトル）名が表示されている。

【0079】

また、このように、それぞれのクラスタ名を、連続的に表示する方法の他に、図9に示すように、それぞれのクラスタ対応のクラスタ名である「用紙」と「カセット」を、それぞれのクラスタ名ごとに改行して縦に並べて表記するようにしてもよい。

【0080】

このように、それぞれのクラスタの名称を縦に並べると、言語的なつながりが気にならなくなり、違和感を与えない効果がある。この実施の形態で用いている「用紙」と「カセット」は、連続して「用紙カセット」としても何等問題ないが、場合によっては、違和感を持つ場合もある。たとえば、これまでの説明とは全く関係のない例として、クラスタマージされた得られた新たなクラスタに含まれ

るそれぞれのクラスタ名が、仮に、「製品」、「使用」、「概要」であったとする。このようなクラスタ名を上述のように、連続して横に一列に並べると「製品仕様概要」となる。これでも意味が全く不明というものではないが、言語的に少し違和感が生じる。このような場合、本来は、言語処理を行って、「製品仕様の概要」というようにすればよいが、そのような言語処理は複雑で時間要する。したがって、このような場合、図9と同様に、「製品」、「使用」、「概要」を1つづつ縦に並べると違和感を与えることがなくなる。また、縦に並べることで、実際に表示したときに、横並び一列での表示に比べ、クラスタマージされたクラスタ名の数が多くても、横方向にむやみに伸びることがないので見易くなるという効果もある。

【0081】

このように、クラスタマージ部2によって得られた関連性を示す値Kが、 $K \geq TH_1$ であって、クラスタマージされて得られた新たにクラスタに含まれるクラスタのクラスタ名をAND形式の表記とし、クラスタ名を横一列に並べた表記の仕方で表示するか、あるいは、各クラスタ対応のクラスタ名称ごとに改行して縦に並べる表記の仕方で表示する。

【0082】

これによって、クラスタマージされて得られた新たなクラスタは、どのようなクラスタがどのような関連性を有して統合されたかということが、そのクラスタマージされた新たなクラスタ名を見るだけでわかる。たとえば、図8や図9の例では、元のクラスタは「用紙」というクラスタと「カセット」というクラスタが統合されてできたクラスタであり、しかも、その関連性はきわめて高く同じような内容の文書を持ったクラスタであるということがわかる。

【0083】

次に、 $TH_1 > K \geq TH$ である場合、すなわち、クラスタマージされて得られた新たなクラスタに含まれる幾つかのクラスタの関連性の度合いは、殆どがオーバーラップするほどでもないが同じ文書を幾つか含んでいるといった場合の処理について説明する。

【0084】

このように、クラスタマージ部2によって得られた関連性を示す値Kが、 $T H 1 > K \geq T H$ である場合には、それぞれのクラスタの名称を、いわゆるOR形式の表記の仕方で行う。

【0085】

たとえば、前述の「製品」、「使用」、「概要」の例で説明すれば、この場合、「製品」、「使用」、「概要」を連続的な表示ではなく、たとえば、「製品・使用・概要」というように、それぞれの名称間に区切りの記号を挿入して表示する。このような区切りの記号がある場合にはOR的な内容であることを予めユーザに報知しておけば、それを見たユーザはそのクラスタマージされて得られた新たなクラスタには、「製品」、「使用」、「概要」といった内容を持った文書が幾つか含まれているというように理解できる。なお、このOR形式の表記の仕方を行う場合、クラスタ名の間に挿入する記号は上述したような「製品・使用・概要」の例に限られるものではなく、たとえば、クラスタ名の間に「/」を挿入して「製品/使用/概要」ようにしてもよい。

【0086】

また、クラスタマージされて得られた新たなクラスタに含まれる幾つかのクラスタの関連性に、 $K \geq T H 1$ と、 $T H 1 > K \geq T H$ が混在するような場合もある。このような場合には、それぞれの関連性の度合いがわかるように、AND形式とOR形式に分けて表記する。

【0087】

さらに、クラスタマージされたそれぞれのクラスタ同志が包含関係にあるような場合もある。たとえば、あるクラスタが「製品」に関するクラスタであり、あるクラスタのクラスタ名が「テレビ」、あるクラスタのクラスタ名が「ラジオ」、あるクラスタのクラスタ名が「ビデオ」であって、これらのクラスタがクラスタマージされたとする。このとき、「テレビ」のクラスタ、「ラジオ」のクラスタ、「ビデオ」のクラスタが「製品」のクラスタに包含されるものであって、しかも、それぞれのクラスタ同志の関連性の度合いが $T H 1 > K \geq T H$ の関係であったとすれば、「製品・(テレビ・ラジオ・ビデオ)」というような表記の仕方

で表示する。これは、「製品」、「テレビ」・「ラジオ」・「ビデオ」はそれぞれがOR的な関係にあり、しかも、「テレビ」・「ラジオ」・「ビデオ」が括弧でくくられていることから、これら「テレビ」・「ラジオ」・「ビデオ」の各クラスタは「製品」に包含されるクラスタであることを意味している。

【0088】

このように、クラスタマージ処理がなされて得られた新たなクラスタのクラスタ名を見るだけで、どのようなクラスタがどの程度の関連性を有して統合されたのかを容易に知ることができる。

【0089】

なお、本発明は以上説明した実施の形態に限定されるものではなく、本発明の要旨を逸脱しない範囲で種々変形実施可能となるものである。たとえば、前述の実施の形態では、図5に示すような分類結果を得るための特徴要素を各文書の表題部から得るようにして、表題部から得られた特徴要素に基づいたクラスタリングを行う例について説明したが、本発明においては、複数の文書をクラスタリングする手法は、特に限定されるものではない。

【0090】

複数の文書をクラスタリングする手法としては、前述の実施の形態で説明した文書の表題部から得られた特徴要素に基づいてクラスタリングを行う例の他に、たとえば、URLアドレス（http://を取り除いた部分）、更新日時（単純な時間または最近1ヶ月以内の更新日時）、ファイルサイズ（webページ本文のバイトサイズなど）を用いてクラスタリングすることもできる。また、これらは、単独で用いてクラスタリングするようにしてもよく、幾つかを組み合わせてもよい。これらのどれを用いるかは、最初にメニューなどで選択項目を選ぶことで可能となる。また、選んだ項目が無い場合には、他の項目を代用する。たとえば、タイトルを選んだ場合、webページにタイトルが無い場合には、URLアドレスを代用する。

【0091】

そして、いずれかの方法によってクラスタリングされたのち、そのクラスタリング結果に対し、前述の実施の形態で説明したような処理、すなわち、それぞれ

のクラスタに含まれる文書の共通性を判断してそれぞれのクラスタ同志を統合するか否かを決めるという処理を施すことによってもクラスタマージを行うことができる。

【0092】

たとえば、URLによってクラスタリングする場合について説明すれば、あるURL（これをURL1とする）のクラスタと、あるURL（これをURL2とする）のクラスタに分類されたとし、URL1のクラスタには文書D1, D2, D3, D4が存在し、URL2のクラスタには文書D2, D3, D4, D5が存在したとする。この場合、これら2つのクラスタには、共通する文書として文書D2, D3, D4が含まれることになり、この共通する文書数と合計の文書数との関係から、URL1のクラスタとURL2のクラスタを統合するか否かを決める。

【0093】

また、クラスタマージするか否かの判断は、前述の実施の形態では、対象となるクラスタに含まれる合計の文書数で共通の文書数を割って得られる割合（%）で表し、その値が予め設定されたしきい値（%）と比較することによって行ったが、これに限られるものではなく、たとえば、共通する文書の個数を数え、その個数とそれぞれのクラスタに含まれる文書数との関係からマージするかしないかを決めるようにすることも可能である。

【0094】

このように、個数によってクラスタマージするか否かを判断する場合、前述したしきい値は個数を用いればよい、たとえば、合計の文書数が10個あって、共通する文書が7個以上であるときにマージするとした場合、前述のしきい値THは、たとえば7個で、TH1をたとえば9個とし、9個以上共通した文書がある場合にはAND形式の表記の仕方での表示を行い、7個または8個の場合はOR形式の表記の仕方での表示を行うというようにもできる。なお、この数値は一例であってこれに限られるものではないことは言うまでもない。これは、前述の実施の形態のなかで説明したしきい値THやTH1の値についても同様のこと�이える。

【0095】

また、前述の実施の形態では、文書D1, D2, …, D7は、それぞれが独立した文書であって、それぞれ独立した文書を分類する場合について説明したが、ある1つの文書を幾つかのコンテンツに分けて、それぞれのコンテンツ（ここでいうコンテンツとは文書の中の意味的なまとまりを指す）を分類する場合にも適用できる。ここで抽出されるコンテンツは、各表題部ごとに切り分けられて得られる文書の中の意味的なまとまりであるとする。

【0096】

たとえば、図2で示した文書D1, D2, …, D7が集まって1つの文書が構成されていると仮定すれば、文書D1, D2, …, D7をそれぞれコンテンツとみなすことができる。これらをコンテンツとすれば、それぞれのコンテンツは、表題部T1, T2, …, T7と本文A1, A2, …, A7から構成されたものとなる。

【0097】

このように、1つの文書を複数のコンテンツに分けて考えた場合、本発明はそれぞれのコンテンツをクラスタリングし、そのクラスタリング結果をクラスタマージする場合にも同様に適応できる。

【0098】

さらに、本発明で用いられるクラスタリング対象文書は、たとえば、汎用の検索サービスで検索された複数の文書をクラスタリング対象文書として考えることもできる。この場合、検索された多数の文書に対してクラスタリング処理を行い、そのクラスタリングされた結果についてクラスタマージ処理を行う。そして、クラスタマージされて得られた新たなクラスタに含まれるそれぞれのクラスタについて前述の実施の形態で説明したよう本発明の処理を行うことで、そのクラスタマージによって得られた新たなクラスタは、もともとどのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかといった内容を容易に知ることができる。

【0099】

また、以上説明した本発明の文書分類処理を行う処理プログラムは、フロッピ

ィディスク、光ディスク、ハードディスクなどの記録媒体に記録させておくことができ、本発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてよい。

【0100】

【発明の効果】

以上説明したように本発明によれば、クラスタマージされて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、それを表示するようにしている。

【0101】

これによって、ユーザは、クラスタマージされる前のクラスタの様子、すなわち、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかといったことを表示内容を見るだけで知ることができる。そして、どのような関連性を有しているかを示す表示の仕方としては、クラスタマージ処理されて得られた新たなクラスタに含まれるクラスタ間の関連性の高さに基づき、クラスタマージ処理されたそれぞれのクラスタのクラスタ名の表示の仕方を変えるようにしている。

【0102】

そのクラスタ名の表示の仕方は、具体的には、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をAND形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さを表す値が予め設定された値未満である場合には、それぞれのクラスタ名をOR形式の表記の仕方で表示させるようにしている。たとえば、関連性の高さがきわめて高い場合には、それぞれのクラスタ名を横一列に連続的に表示したり、それぞれのクラスタ名を1つづつ縦に並べて表示し、関連性の高さがそれほどでもない場合には、それぞれのクラスタ名の間に区切り記号を挿入するなどして表示する。ユーザはこのような表示を見ることで、統合される前のそれぞれのクラスタがどのようなクラスタであって、それぞれのクラスタ同志の関連性がどの程度であるかなどを知ることができる。

【0103】

また、あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することも可能であり、包含関係であることを繁雑なイメージを使わないでもわかりやすく表示できる。

【0104】

このように本発明によれば、複数の文書を分類処理することによって生成された多数のクラスタに対し、クラスタマージ処理を施すことにより関連性の高いクラスタ同志を統合してまとめることができるので、たとえば、検索サービスなどで検索された多数の文書に対し、本発明を適用することにより、検索要求を出したユーザに対し、検索結果を簡潔化した見易いクラスタリング結果として提示することができる。これによって、ユーザは自分の欲しい情報を効率よく探すことができ、従来にはない検索サービスが実現できる。

【図面の簡単な説明】

【図1】

本発明の文書分類装置の実施の形態を説明するブロック図である。

【図2】

本発明の実施の形態を説明するための複数の文書例を示す図である。

【図3】

本発明が行う文書分類処理の処理手順を概略的に説明するフローチャートである。

【図4】

特徴要素とそれぞれの文書との関係を示す特徴テーブル内容の一例を示す図である。

【図5】

図4に示す特徴テーブルに基づいて文書を分類した分類結果を示す図である。

【図6】

2つのクラスタ間でのクラスタマージ処理を説明する図であり、それぞれのクラスタに含まれる文書例を示す図である。

【図7】

図5の分類結果についてクラスタマージ処理した結果を示す図である。

【図8】

クラスタマージされて得られた新たなクラスタに含まれるそれぞれのクラスタのクラスタ名をAND形式（横一列に並べた場合）の表記の仕方で表示した例を示す分類結果例を示す図である。

【図9】

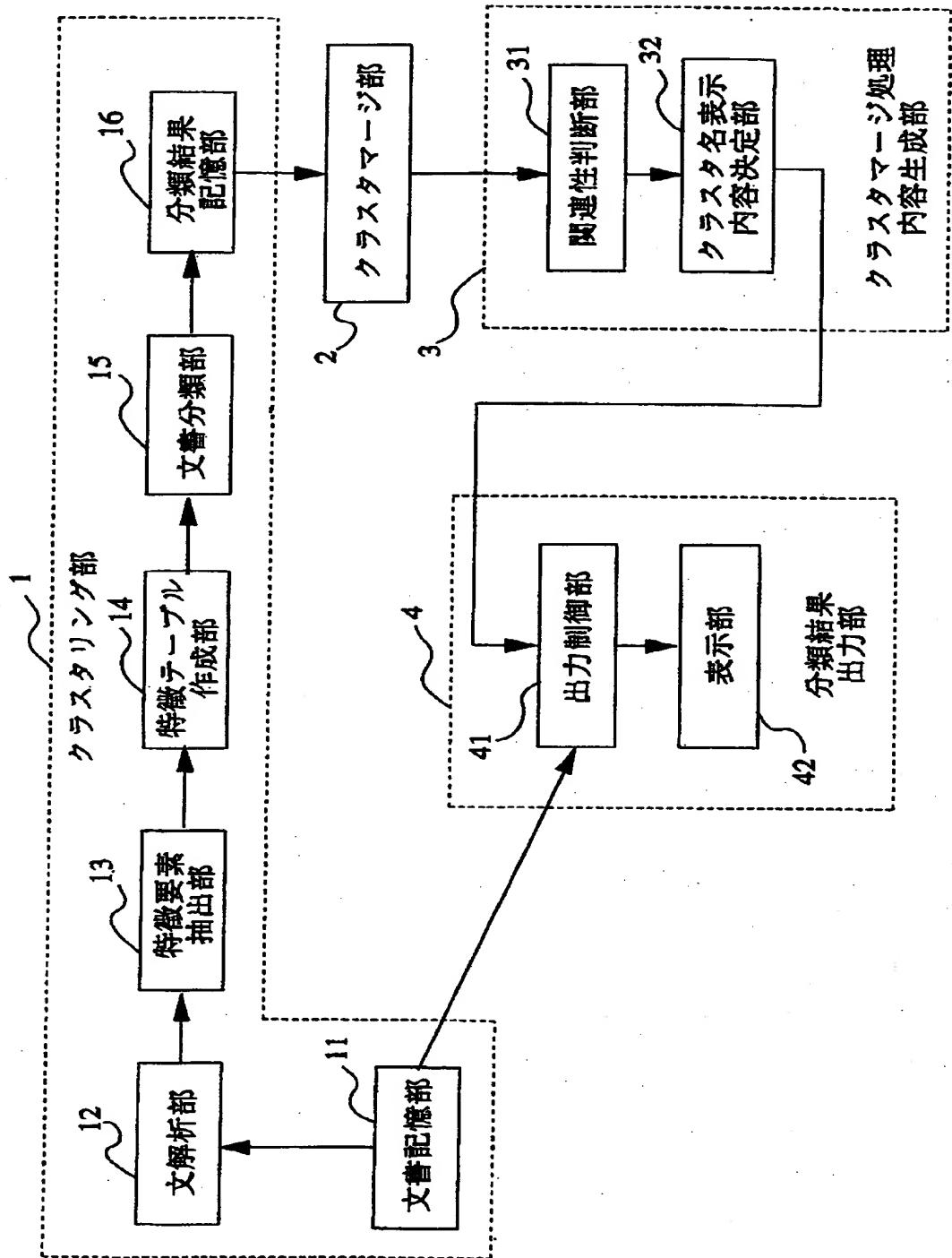
クラスタマージされて得られた新たなクラスタに含まれるそれぞれのクラスタのクラスタ名をAND形式（クラスタ名を1つづつ縦に並べた場合）の表記の仕方で表示した例を示す分類結果例を示す図である。

【符号の説明】

- 1 クラスタリング部
- 2 クラスタマージ部
- 3 クラスタマージ処理内容生成部
- 4 分類結果出力部
 - 1 1 文書記憶部
 - 1 2 文解析部
 - 1 3 特徴要素抽出部
 - 1 4 特徴テーブル作成部
 - 1 5 文書分類部
 - 1 6 分類結果記憶部
- 3 1 関連性判断部
- 3 2 クラスタ名表示内容決定部
- 4 1 出力制御部
- 4 2 表示部
 - A 1, A 2, …, A 7 本文
 - D 1, D 2, …, D 7 文書
 - T 1, T 2, …, T 7 表題部

【書類名】 図面

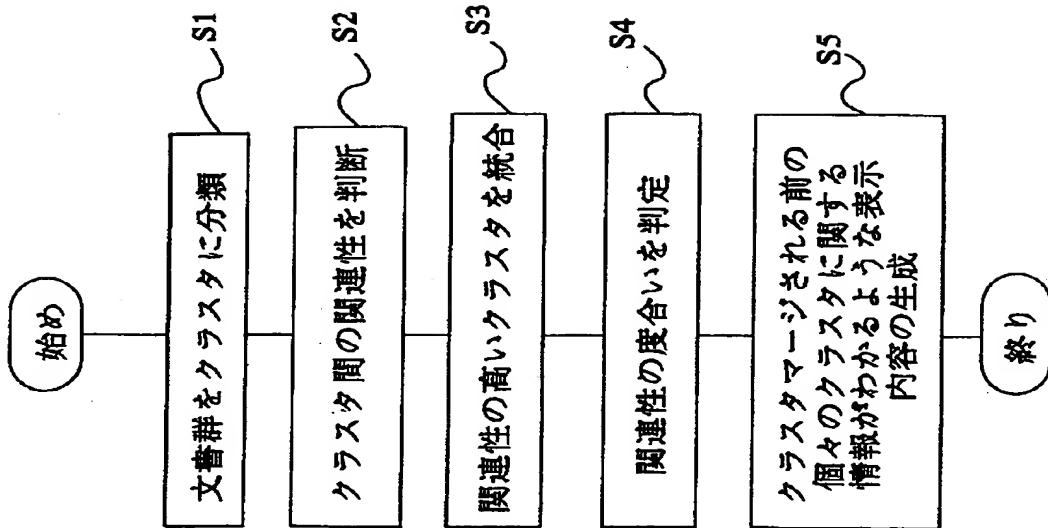
【図1】



【図2】

- T1 *用紙カセットについて
A1 標準装備のユニバーサル用紙カセットはオプションのA4専用の大容量用紙カセットに取り替えることが可能である。標準ではトレイに200枚の用紙をセットすることが・・・
- D1
- T2 *レーザプリンタのメモリの増設について
A2 レーザプリンタのメモリを増設することでパソコンの開放時間を早めたり、・・・することが可能になる。ただし、どの程度・・・
- D2
- T3 *オプションインターフェースカードの増設について
A3 オプションのインターフェースカードを使用して、ネットワーク上にプリンタをダイレクトに接続して使用することができる。そして、・・・
- D3
- T4 *用紙設定で「トレイ」「カセット」「ジドウ」の切替えについて
A4 様々なアプリケーションから印刷する際、給紙装置や用紙サイズの設定をする必要がある。用紙カセットには用紙ガイドタリップが・・・
- D4
- T5 *プリンタにフォントを追加するためにハードディスクを増設する場合について
A5 フォントを追加する場合は、・・・する方法がある。また、オプションのフォントROMボードを装着する場合は、・・・することができる。
- D5
- T6 *印刷後における出力された用紙の汚れについて
A6 用紙の端や裏面に黒い汚れがつく場合には、プリンタ本体定着器ローラのクリーニングを実施する。プリンタ本体のパネルから「クリーニング印刷」の設定とし、A4用紙にクリーニング用紙を印刷する。次に・・・
- D6
- T7 *用紙カセットの増設について
A7 オプションのダブルカセットユニットもしくはA4専用の大容量用紙カセットユニットが2つまで装着可能である。・・・
- D7

【図3】



【図4】

特徴要素	文書 D1	文書 D2	文書 D3	文書 D4	文書 D5	文書 D6	文書 D7
用紙	1			1		1	1
カセット	1			1			1
増設		1	1		1		1

【図5】

特徴要素	クラスタ
用紙	D1, D4, D6, D7
カセット	D1, D4, D7
増設	D2, D3, D5, D7

【図6】

クラスC1	D1, D2, D3, D4, D8
クラスC2	D3, D4, D5, D6, D7, D8

【図7】

特徴要素	クラス
用紙+カセット	D1, D4, D6, D7
増設	D2, D3, D5, D7

【図8】

クラス名	文書のタイトル
用紙カセット	<ul style="list-style-type: none">*用紙カセットについて*用紙設定で「トレイ」「カセット」「ジドウ」「ジドウ」の切替えについて*印刷後における出力された用紙の汚れについて*用紙カセットの増設について

【図9】

クラスタ名	文書のタイトル
用紙カセット	<ul style="list-style-type: none">* 用紙カセットについて* 用紙設定で「トレイ」「カセット」「ジドウ」の切替について* 印刷後における出力された用紙の汚れについて* 用紙カセットの増設について

【書類名】 要約書

【要約】

【課題】 多数の文書をそれぞれの文書に存在する特徴要素に基づいてクラスタに分類する場合、多数のクラスタを整理しわかりやすくして出力する。

【解決手段】 それぞれの文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部1と、得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部2と、クラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ内容生成部3と、その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力手段4とを有する。

【選択図】 図1

出願人履歴情報

識別番号 [000002369]

1. 変更年月日 1990年 8月20日

[変更理由] 新規登録

住 所 東京都新宿区西新宿2丁目4番1号

氏 名 セイコーエプソン株式会社